САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
КАФЕДРА ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Лавелина Елизавета Сергеевна

Магистерская диссертация

Вебометрические методы в исследовании характеристик веб-пространств крупных организаций

Направление 020402
Фундаментальная информатика и информационные технологии
Магистерская программа «Технологии баз данных»

Научный руководитель,
доктор технических. наук,
профессор кафедры
технологии программирования
Печников А.А.

Санкт-Петербург
2019

Содержание

Введение

С момента своего создания Интернет произвел революцию в повседневной жизни, предоставляя пользователям доступ к огромному количеству информации. Сегодня в Вебе существует множество веб-сайтов, взаимодействие между которыми осуществляется при помощи гиперссылок. Несмотря на огромное количество, веб-сайты образуют вполне упорядоченную систему. Например, наличие гиперссылок может приводить к увеличению количества переходов между сайтами, а значит и к росту числа посетителей.

Подобным образом некоторые веб-сайты могут искусственно увеличивать свою популярность путем обмена ссылками. Одним из примеров могут являться так называемые “малые Интернет-сообщества” — множество веб-сайтов крупных предприятий, университетов, научных центров и т. д., связанных при помощи гиперссылок. Данные сообщества содержат небольшое количество участников, что объясняет их название. Участники таких сообществ могут согласовывать свои действия для увеличения ссылочной популярности, и, как следствие, рейтинга в выдаче поисковых систем.

Исследованиями сети Интернет занимается вебометрика — одно из научных направлений, в рамках которого, в частности, изучается взаимодействие веб-сайтов, их структура, а также исследуются их количественные характеристики. Данные исследования помогают глубже понять связи между различными сообществами, а также выявить закономерности в расставлении ссылок между веб-сайтами. Например, они помогают определить, насколько организация следит за своими веб-сайтами и тенденциями развития сети Интернет.

Постановка задачи

Цель работы заключается в исследовании веб-пространства коммерческой организации и выявления закономерностей распределения внешних гиперссылок.

Для достижения выше поставленной цели необходимо было решить следующие задачи:

- сбор внешних гиперссылок выделенного Интернет-сообщества и создание базы данных внешних гиперссылок
- построение и анализ веб-графа
- исследование веб-ссылок выделенного Интернет-сообщества для выявления признаков согласованного поведения

Обзор литературы

Интернет-пространство можно рассматривать с различных точек зрения для изучения уникальных онлайн-явлений, а также оффлайн-явлений, которые отражаются в сети. Анализ можно проводить на основе содержания сайта, структуры веба или поведения пользователя [1]. Основным компонентом, доступным для анализа, является гиперссылка.

Вебометрические методы основаны на исследовании информации, которая содержится в гиперссылках, соединяющих различные веб-страницы в Интернете. Термин «вебометрика» был впервые введен Томасом Алминдом и Петером Ингверсеном в 1997 году [2]. Большинство исследований в этой области были сосредоточены на академических и научных веб-пространствах [3], а также на социальных сетях [4, 5]. Однако эта методология в равной степени применима к коммерческим сайтам, которые более распространены в Интернете [6].

Основными направлениями вебометрики являются анализ содержания веб-страниц (webpage content analysis) [7], анализ использования сети (web usage analysis) [8] и анализ структуры веб-ссылок (weblink structure analysis) [9]. Анализ содержания веб-страниц сфокусирован на анализе содержимого веб-страниц: текста, изображений, видео и др. Данное направление тесно связано с дата майнингом [10], так как многие технологии могут быть применены к анализу содержимого веб-страниц, в частности NLP (Natural Language Processing) и IR (Information retrieval) [11]. Анализ использования сети применяет технологии, которые позволяют предсказать поведение пользователей сети Интернет, а также выделить паттерны использования данных в Вебе для лучшего понимания нужд пользователей. Анализ структуры веб-ссылок позволяет понять как связаны различные веб-страницы, а также по каким принципам были построены веб-сайты.

Вебометрический анализ основан на данных, собранных в Интернете. В частности, источниками данных для исследований структуры веб-ссылок могут выступать коммерческие поисковые системы, а также веб-краулеры — специальные программы для сбора данных о Вебе [12]. Сегодня большинство крупных поисковых систем обладают возможностью расширенного поиска, в частности такая возможность есть у Google [13], Yandex [14], Yahoo [15], Bing [16], DuckDuckGo [17]. Что касается веб-краулеров, наиболее известными являются SocSciBot и LexiURL, разработанные профессором Майклом Терваллом [4]. Также существует большое количество других программ-краулеров, с помощью которых можно собирать данные для конкретных исследовательских задач [18].

Исследование веб-ссылок тесно связано с другой проблемой — ранжирования результатов поиска. Большинство алгоритмов ранжирования так или иначе учитывают ссылочную популярность ресурса при поисковой выдаче [19]. Одной из проблем таких алгоритмов является возможность искусственного увеличения ссылочной популярности, а, следовательно, и места сайта на странице выдачи результатов поиска, путем обмена веб-ссылками [20]. Конечно, не всегда обмен ссылками производится намеренно и является накруткой — многие крупные веб-сайты, имеющие схожую тематику, имеют ссылки друг на друга, что нельзя считать договорными действиями [21]. Для подтверждения теории о согласованном поведении Интернет-сообществ можно использовать математические модели, которые имеют соответствующую целевую функцию [22].

Глава 1. Теоретическая часть

В данной главе излагаются основные понятия предметной области, а также описываются математические модели согласованного поведения сообществ.

1.1 Основные понятия и определения

Введем некоторые формальные определения, которые будут использованы для последующего изложения.

Веб-сайт — совокупность html-страниц, связанных гиперссылками, которые принадлежат одному домену. Веб-пространство организации — это множество веб-сайтов, принадлежащих одной организации, которые связаны с помощью гиперссылок. У веб-пространства можно выделить его главный сайт — так называемый «головной сайт», чаще всего это официальный сайт организации. Различают два типа гиперссылок: внешние и внутренние. Внутренние гиперссылки — гиперссылки, связывающие html-страницы одного веб-сайта, которые принадлежат одному домену. Внешние гиперссылки — гиперссылки, которые ссылаются на веб-сайты, не принадлежащие текущему домену.

Веб-краулер — программа для сбора данных о Вебе путем последовательного перехода по веб-страницам [23]. В общем случае на вход краулеру подается список URL-адресов веб-сайтов для сканирования. На каждой итерации он извлекает один из сайтов, находит все содержащиеся на нем гиперссылки и добавляет их к списку сканирования, если они еще не были посещены. Сканирование осуществляется, пока список еще не посещенных страниц не будет пуст, либо пока не будет достигнута заданная глубина сканирования. Глубина сканирования — уровень веб-страницы, до которого должен быть просканирован сайт. Уровень веб-страницы определяется следующим образом: начальная страница, определяемая по доменному имени

сайта, имеет уровень 0. Уровень любой другой страницы — это минимальное количество внутренних гиперссылок, ведущих от начальной страницы к данной.

При помощи краулера извлекаются данные для последующего анализа веб-пространства, в том числе их можно использовать для построения веб-графа. Веб-граф — это ориентированный граф, вершинами которого являются html-страницы, а ребрами — гиперссылки, связывающие данные вершины [24].

Важной характеристикой Интернет-ресурсов также является их ссылочная популярность, так как она используется при расчете показателей PageRank, HITS и т. д., результаты которых учитываются при ранжировании ссылок в поисковых системах. Можно сказать, что значимость Интернет-ресурса в той или иной мере зависит от количества ссылок на данный ресурс.

1.2 Анализ структуры веб-ссылок

В вебе существует около одного триллиона веб-страниц с уникальным URL, которые проиндексированы поисковыми системами. Реальное количество может быть намного больше, так как постоянно появляются новые веб-страницы, которые еще не добавлены в индекс. С момента появления сети Интернет количество веб-страниц растет с огромной скоростью, как и количество пользователей. Поэтому важно понимать и анализировать структуру Веба для эффективного поиска информации.

Веб содержит множество веб-сайтов, которые не имеют единой структуры как внутренних, так и внешних ссылок. Несмотря на то что гиперссылки предназначены для навигации между веб-страницами, они также содержат скрытую информацию, которая может быть использована для исследования Веба. Основная цель анализа веб-ссылок — понимание внутренней организации сети, извлечение скрытых характеристик веб-сайтов, а также определение сходств и различий между ними.

Существует два основных вида анализа веб-ссылок: оценка влияния ссылок (link impact assessments) и анализ связей между гиперссылками (link relationship mappings). Для оценки влияния ссылок выделяется некоторое множество веб-сайтов, для которых сравнивается число ведущих на них внешних гиперссылок. Это позволяет понять, насколько высокую ссылочную популярность имеют веб-сайты из множества. Анализ связей между гиперссылками также производится для некоторого набора веб-сайтов. Для иллюстрации отношений между веб-сайтами строится веб-граф, который может использоваться для анализа связей между веб-сайтами, а также для выявления паттернов распределения гиперссылок.

1.3 Построение математической модели

Исследование веб-пространств может осуществляться с помощью математических моделей. В данной работе рассматривается одна из моделей согласованного поведения Интернет-сообществ.

Рассмотрим малое Интернет-сообщество, которое имеет следующие характеристики:

n — количество участников сообщества

c_i — значимость i-го участника, $c_i \geq 0, i = \overline{1,n}$.

m_i — количество прямых ссылок от i-го участника на других участников сообщества, $m_i \geq 0, \forall\, i = \overline{1,n}$.

$X(x_{ij}),\ i,j = \overline{1,n}$ — матрица ссылок, где $x_{ij} = 1$, если есть ссылка от i-го участника к j-му, и $x_{ij} = 0$, если ссылки нет.

Введем систему следующих ограничений:

- $x_{ii} = 0, i = \overline{1,n}$ — ресурс не должен содержать ссылки на себя;
- $x_{ij} = 0,\ i = \overline{1,n},\ j = \overline{1,n}$ — ссылки либо существуют, либо не существуют;

- $\sum_{j=1}^{n} x_{ij} \leq m_i,\ i = \overline{1,n}$ — количество исходящих ссылок ограничено;
- $\sum_{j=1}^{n} x_{ij} \geq 1, i = \overline{1,n}$ — участниками сообщества являются ресурсы, от которых исходит хотя бы одна ссылка, т. е. они делятся своей значимостью с другими участниками.

Для определения функции значимости используем подход, основанный на линейном представлении функции приращения, или, другими словами, на решении системы линейных уравнений. Основными предположениями, на основе которых вычисляются данные функции, являются:

- чем больше ссылок на ресурс, тем он более «значимый»;
- значимость ресурса j возрастает, если увеличивается значимость ресурса i, при $x_{ij} = 1$;
- приращение значимости ресурса j будет уменьшаться с увеличением числа исходящих ссылок от ресурса i, при $x_{ij} = 1$.

Предположим, что изменение значимости j-го участника сообщества представлено следующей формулой:

$$\forall j = \overline{1,n} : \hat{c}_j = c_j + \sum_{i=1}^{n} x_{ij} \cdot c_i \cdot \alpha_i$$

где α_i — коэффициент, показывающий, что при $x_{ij} = 1$ (то есть при установлении ссылки с i-го участника на j-го) значимость j-го участника возрастает на некоторую часть значимости i-го участника.

Пусть $F(X)$ — функция, характеризующая некоторый интегральный показатель значимости всех участников сообщества, зависящая от матрицы X, то есть от того, как расставлены ссылки между участниками. Тогда задача заключается в нахождении такой расстановки прямых ссылок, при которой увеличивается значимость всех сайтов сообщества, то есть в нахождении такой

матрицы X, которая удовлетворяет заданным ограничениям и дает оптимальное значение целевой функции $F(X) \to \underset{x_{ij}}{opt}$.

В качестве целевой функции возьмем функцию среднеквадратичного отклонения [25]. Оптимизационная задача принимает следующий вид при ограничениях, введенных выше:

$$F1(X) = \sum_{j=1}^{n} \left(\frac{\sum_{k=1}^{n} \hat{c}_k}{n} - \hat{c}_j \right)^2 \to \min_{x_{ij}}$$

Согласованные действия участников сообщества при данной оптимизационной функции можно сформулировать следующим образом: расстановка прямых ссылок внутри сообщества должна привести к минимальному отклонению полученных значимостей каждого участника от нового среднего значения по всему сообществу.

В качестве целевой функции также можно рассматривать линейную функцию вида [26]:

$$F2(X) = \sum_{j=1}^{n} \hat{c}_j \cdot \lambda_j \to \max_{x_{ij}},$$

где коэффициенты $0 < \lambda_j \leq 1$ зависят от значения c_j.

В этом случае согласованные действия участников сообщества можно сформулировать как: суммарный прирост значимости внутри сообщества должен быть максимальным, за счет увеличения наименее низких значимостей участников.

Определим коэффициент $\lambda_j = \frac{1}{c_j}$, то есть λ_j обратно пропорционален значимости участника. Тогда целевая функция принимает вид:

$$F2(X) = \sum_{i=1}^{n} \sum_{j=1}^{n} x_{ij} \cdot \frac{c_i}{c_j} \cdot \alpha_i \to \max_{x_{ij}}$$

Пусть $K = \sum_{i=1}^{n} \frac{\hat{c_i}}{n}$ — среднее значение получаемых значимостей. В таком случае, если значимость некоторого участника системы изначально больше значения K, получаемого в процессе решения задачи, то на него не нужно устанавливать ссылки: $\sum_{j=1}^{n} x_{ij} = 0, \forall j : c_j \geq K$.

В данном исследовании можно заменить два последних ограничения на ограничение вида:

$$\sum_{j=1}^{n} x_{ij} = m_i \, , i = \overline{1,n}$$

Данная замена может быть осуществлена, так как в том случае, если поведение участников является согласованным, то они должны договориться о количестве исходящих прямых ссылок $m_i > 0$ каждого участника. Тогда количество исходящих ссылок на других участников сообщества является константой.

Таким образом, *Модель 1* согласованного поведения участников сообщества принимает следующий вид:

$$F(X) = \sum_{j=1}^{n} \left(K - (c_j + \sum_{i=1}^{n} \frac{x_{ij}}{L_i} \cdot c_i) \right)^2 \to \min_{x_{ij}}$$

при ограничениях:

$$x_{ii} = 0, \; i = \overline{1,n},$$

$$x_{ii} = 0, \; i = \overline{1,n}, \; j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = m_i \, , \; i = \overline{1,n}.$$

где функция приращения значимости:

$$\forall j = \overline{1,n} : \hat{c}_j = c_j + \sum_{i=1}^{n} x_{ij} \cdot c_i \cdot \alpha_i$$

Модель 2:

$$F2(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij} \cdot \frac{c_i}{c_j} \cdot \alpha_i \to \max_{x_{ij}}$$

при ограничениях:

$$x_{ii} = 0,\ i = \overline{1,n},$$

$$x_{ii} = 0,\ i = \overline{1,n},\ j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = 0, \forall j : c_j \geq K,$$

$$\sum_{j=1}^{n} x_{ij} = m_i, i = \overline{1,n},$$

где функция приращения значимости:

$$\forall j = \overline{1,n} : \hat{c}_j = c_j + \sum_{i=1}^{n} x_{ij} \cdot c_i \cdot \alpha_i$$

1.4 Уточнение математической модели

Внесем некоторые уточнения в математическую модель согласованного поведения участников сообщества. Будем полагать, что на фиксированном временном интервале новые ссылки на участников сообщества не добавляются, а уже существующие не исчезают. Следовательно, изменение функции значимости зависит только от расстановки ссылок между участниками сообщества.

Введем следующие обозначения:

- $L_i,\ i = \overline{1,n}$ — общее количество исходящих ссылок от i-го участника сообщества;
- $\hat{L}_i = L_i - m_i,\ i = \overline{1,n}$ — количество ссылок, исходящих от i-го участника сообщества без учета ссылок на других участников сообщества.

Поскольку увеличение числа исходящих ссылок уменьшает приращение значимости ресурса, уточним вид коэффициента α_i. Выразим α_iкак:

$$\alpha_i = \frac{\beta}{L_i}$$

где β — параметр конкретного алгоритма вычисления значимости, зависящий от поисковой машины. Получить точное значение коэффициента β не представляется возможным, так как, во-первых, данное значение является коммерческой тайной любой поисковой системы, во-вторых, алгоритмы ранжирования поисковых систем постоянно корректируются для получения более точных результатов, а, следовательно, изменяется значение параметра β. Попытки оценить значение β приводят к достаточно большому разбросу значений в пределах от 0,3 до 2,83, поэтому в данной работе будем считать $\beta = 1$.

Тогда изменение значимости j-го участника будет вычисляться следующим образом:

$$\forall j = \overline{1,n} : \hat{c}_j = c_j + \sum_{i=1}^{n} x_{ij} \cdot c_i$$

Выразим коэффициент Kс учетом приведенных уточнений:

$$K = \left(\sum_{j=1}^{n} c_j + \sum_{i=1}^{n} \frac{m_i \cdot c_i}{L_i} \right) / n$$

Благодаря тому что полученное значение Kне зависит от матрицы X, вычисление целевых функций $F1(X)$ и $F2(X)$ значительно упрощается.

С учетом всех уточнений, сделанных выше, полученные математические модели выглядят следующим образом:

Модель 1:

$$F(X) = \sum_{j=1}^{n}\left(K - (c_j + \sum_{i=1}^{n}\frac{x_{ij}}{L_i}\cdot c_i)\right)^2 \to \min_{x_{ij}}$$

при ограничениях:

$$x_{ii} = 0,\ i = \overline{1,n},$$

$$x_{ij} = 0{,}1,\ i = \overline{1,n},\ j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = m_i,\ i = \overline{1,n}.$$

Модель 2:

$$F2(X) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}\cdot\frac{c_i}{c_j}\cdot\alpha_i \to \max_{x_{ij}}$$

при ограничениях:

$$x_{ii} = 0,\ i = \overline{1,n},$$

$$x_{ij} = 0{,}1,\ i = \overline{1,n},\ j = \overline{1,n},$$

$$\sum_{j=1}^{n} x_{ij} = 0, \forall j : c_j \geq K,$$

$$\sum_{j=1}^{n} x_{ij} = m_i, i = \overline{1,n}.$$

В качестве критерия, который позволит определить степень согласованности поведения сообщества, используется отклонение значения функционалов $F1(X)$ и $F2(X)$, вычисленных для реальной матрицы сообщества, от значений этих же функционалов на матрице оптимального решения. В сущности, сравнивается отклонение функционалов $F1(X)$ и $F2(X)$, вычисленных на реальной матрице $X^{real} = \left(x_{i,j}^{real}\right)$, от оптимальных значений, вычисленных на матрице X^{opt}, которая отличается от X^{real}с точностью до перестановки единиц в строках.

При вычислении функционалов $F1(X^{real})$ и $F2(X^{real})$ используются значения $\hat{c}_j,\ i = \overline{1,n}$, которые получены в результате согласованных действий

участников сообщества. Для вычисления $F1(X^{opt})$ и $F2(X^{opt})$ необходимо использовать исходные значения $c_i, i = \overline{1,n}$, которые могут быть найдены из системы линейных уравнений:

$$\begin{cases} c_1 + \sum_{i=1}^{n} \frac{x_{i1}}{L_i} \cdot c_i = \hat{c}_1 \\ \dots \\ c_k + \sum_{i=1}^{n} \frac{x_{ik}}{L_i} \cdot c_i = \hat{c}_k \\ \dots \\ c_n + \sum_{i=1}^{n} \frac{x_{in}}{L_i} \cdot c_i = \hat{c}_n \end{cases}$$

Глава 2. Построение веб-графа веб-пространства коммерческой организации

В данной главе описаны основные программные средства, используемые для сбора и обработки информации, а также рассмотрено применение теоретико-графовых методов для исследования выделенного веб-сообщества.

2.1 Инструменты

Для сбора и анализа гиперссылок с целью построения веб-графа, а также для последующего исследования веб-графа использовались следующие вебометрические инструменты:

- программный компонент для поиска и сбора внешних гиперссылок
- база данных для хранения внешних гиперссылок
- открытая платформа для визуализации графов

Для сбора данных было решено использовать готовую программу-краулер. Выбор краулера осуществлялся, основываясь на следующих требованиях к программе:

- В качестве исходных данных передается доменное имя головного сайта исследуемого веб-пространства крупной организации и максимальная глубина сканирования каждого сайта веб-пространства.
- Обход каждого сайта, начиная с главной заданной страницы, осуществляется «в ширину» по внутренним гиперссылкам.
- Объекты сканирования – только html-страницы.
- Сканирование осуществляется до тех пор, пока не будет достигнута заданная глубина сканирования, либо список страниц, которые необходимо посетить будет пуст.

- Краулер не должен делать запросы слишком большое количество раз в секунду — должен быть «вежливым».

В итоге был выбран Crawler4j [27] — параллельный, кроссплатформенный краулер, реализованный на Java. С его помощью можно легко организовать многопоточный краулинг, что значительно ускоряет время сбора ссылок.

Для хранения собранных ссылок было решено использовать реляционную базу данных PostgreSQL [28]. Преимуществом данной базы данных является то, что это бесплатная кроссплатформенная Open Source система, которая предоставляет практически все возможности, которые есть в других базах данных (коммерческих или Open Source). Также она достаточно надежна и имеет хорошие характеристики по производительности.

Для построения графа была выбрана программа Gephi [29] — открытая платформа для визуализации графов. Она содержит ряд стандартных возможностей, удобных для исследования графа, таких как разбиение множества ссылок на кластеры и оценок значимости вершин по алгоритму ссылочного ранжирования.

2.2 Выделение сообщества

В данной работе в качестве объекта исследования было выбрано одно из малых Интернет-сообществ, а именно группа сайтов коммерческой организации «Газпром». Всего данное сообщество включает 117 сайтов. Предполагается, что такой объект исследования является достаточно типичным сообществом, и поэтому методы исследований могут быть перенесены на другие крупные коммерческие организации.

2.3 Сбор данных

В работе был применен следующий алгоритм сбора данных для построения веб-пространства коммерческой организации:

- Составляется список всех сайтов, принадлежащих организации.
- Каждый сайт из списка сканируется при помощи веб-краулера.
- Результаты работы краулера вносятся в базу данных.

Список сайтов, подлежащих сканированию содержал 117 веб-сайтов веб-пространства компании Газпром (табл. 1). В данном случае сайты, являющиеся поддоменами головного сайта «www.gazprom.ru», входящие в список, считались самостоятельными и сканировались отдельно.

Доменное имя веб-сайта	Официальное название веб-сайта
http://gazprom.ru/	Газпром
http://gazprom.com/	Gazprom
http://eco-gas.ru/	EcoGas
http://www.metan.by/	EcoGas в Белоруссии
http://www.gazpromvacancy.ru/	Вакансии
http://ecogas-auto.ru/	Газомоторная техника
http://vbashkortostane.gazprom.ru/	«Газпром» в Башкортостане
http://nakubani.gazprom.ru/	«Газпром» на Кубани
http://gazprompolus.ru/	«Газпром» на Южномполюсе
http://gazpromvideo.ru/	Газпром видео

Табл. 1 Список сайтов компании Газпром (первые 10 строк).

При сканировании каждого сайта из списка производились следующие действия для отбора внешних гиперссылок:

- Все ссылки проверяются с помощью следующего паттерна: «css | js | bmp | gif | jpeg | png | tiff | ico | nef | raw | mid | mp2 | mp3 | mp4 | wav | wma | flv | mpeg | avi | mov | mpeg | ram | m4v | wmv | rm | smil | pdf | doc | docx | pub | xls | xlsx | vsd | ppt | pptx | swf | zip | rar | gz | bz2 | 7z | bin | xml | txt | java | c | cpp | exe». Если они подходят по него, то больше не рассматриваются, так как рассматриваются только ссылки на html-страницы.
- Проверяется, является ли гиперссылка внешней, и, если это не так, то она также отбрасывается.
- Полученные внешние ссылки усекаются до доменного имени.
- Суммируется количество ссылок, ведущих на один и тот же веб-сайт.

При сборе ссылок, после усечения ссылки до доменного имени, были выявлены следующие особенности: ссылки ведущие на один и тот же сайт могли иметь разное написание, а именно, часть ссылок начиналась с www, другая — нет. Например, при краулинге сайта «www.gazprom.com» было найдено и ссылки, ведущие на «www.gazprom.ru», и ссылки на «gazprom.ru». Для устранения неоднозначности написания ссылки сначала записывались в промежуточную таблицу, которая имела следующий вид:

- link_id — уникальный номер записи
- seed — доменное имя ссылки-источника, откуда была собрана внешняя гиперссылка
- link_path — домен ссылки-приемника, сама внешняя гиперссылка
- page_amount — количество гиперссылок, не содержащих «www.»
- page_amount_www — количество гиперссылок, начинающихся с «www.»

Затем на основе промежуточной таблицы при помощи хранимой процедуры была сформирована итоговая таблица, имеющая следующий вид:

- link_id — уникальный номер записи

- seed — доменное имя ссылки-источника, откуда была собрана внешняя гиперссылка
- link_path — домен ссылки-приемника, сама внешняя гиперссылка
- page_amount — количество гиперссылок

В данном случае производилось полное сканирование сайтов. Часть таблицы, которая была получена в результате работы краулера по сбору внешних гиперссылок, приведена в таблице 2.

link_id	seed	link_path	page_amount
1	www.gazprom.ru	www.mosenergo-museum.ru	16185
2	www.gazprom.ru	www.artlebedev.ru	2
3	www.gazprom.ru	turkstream.info	16186
4	www.gazprom.ru	www.kommersant.ru	3
5	www.gazpromspartakiada.ru	www.instagram.com	1896
6	www.gazpromspartakiada.ru	gazprom.ru	893
7	turkstream.info	www.artlebedev.com	4
8	turkstream.info	www.south-stream-transport.com	2
9	vostokgazprom.gazprom.ru	polyanaski.ru	267
10	vostokgazprom.gazprom.ru	twitter.com	266

Табл. 2 Часть полученной итоговой таблицы

2.4 Построение веб-графа

Для построения веб-графа на вход Gephi подавался csv-файл со списком вершин, полученный из итоговой таблицы. Граф содержит 2817 вершин и 9872 ребра. Изображение полученного графа представлено на рис. 1. Красным цветом выделены сайты «www.gazprom.ru» и «www.gazprom.com». Голубым окрашены

сайты, которые принадлежат веб-пространству «Газпрома», черным — остальные внешние веб-сайты.

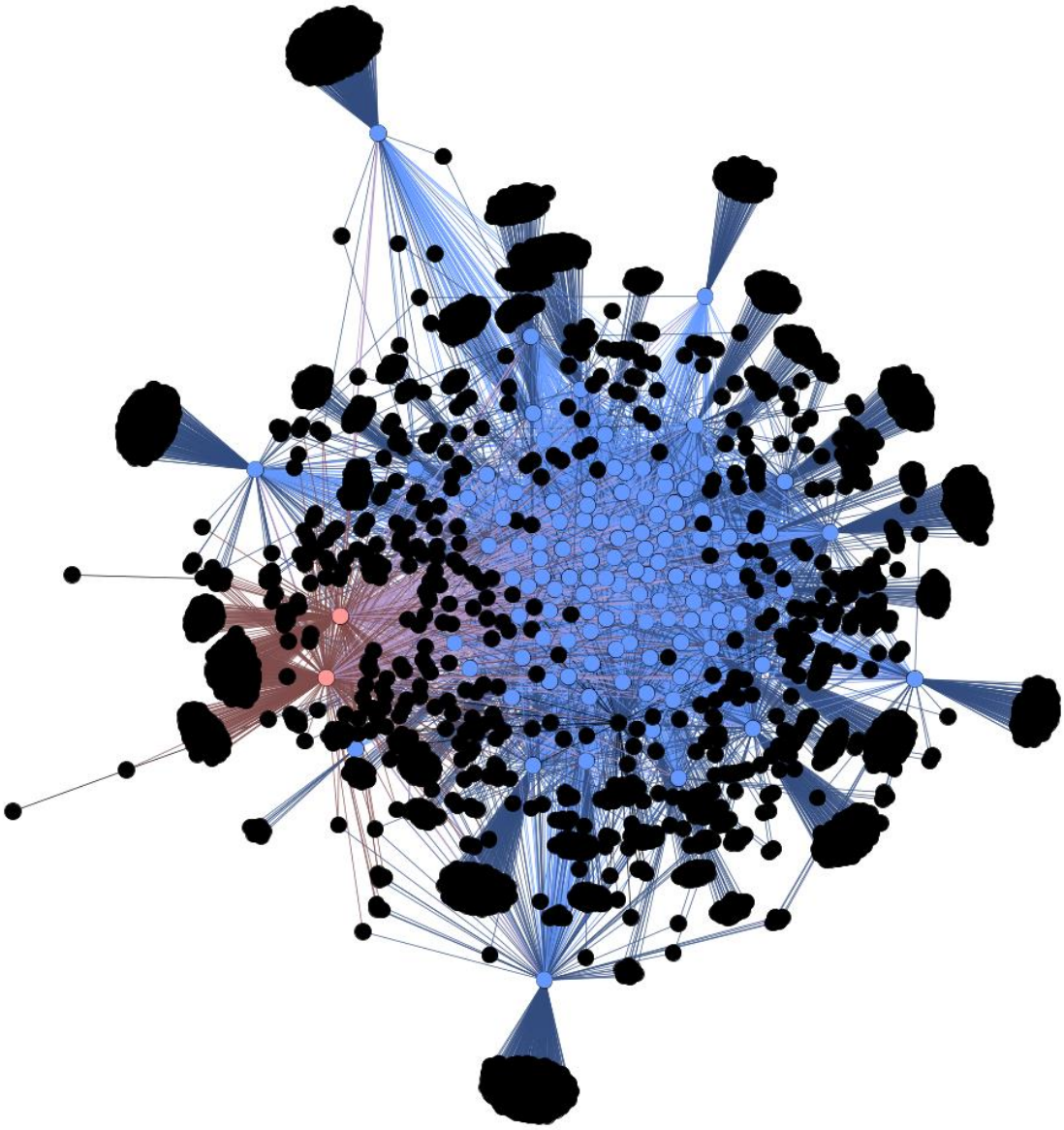

Рис. 1 Веб-граф внешних гиперссылок и веб-пространства компании «Газпром»

Проанализировав полученный веб-граф, можно увидеть, что сайты веб-пространства «Газпром» имеют множество ссылок друг на друга. Для более детального изучения был построен веб-граф только для веб-сайтов, входящих в список сайтов компании «Газпром» (рис. 2). Данный граф содержит 117 вершин

и 4922 ребра. Красным выделены сайты «www.gazprom.ru» и «www.gazprom.com».

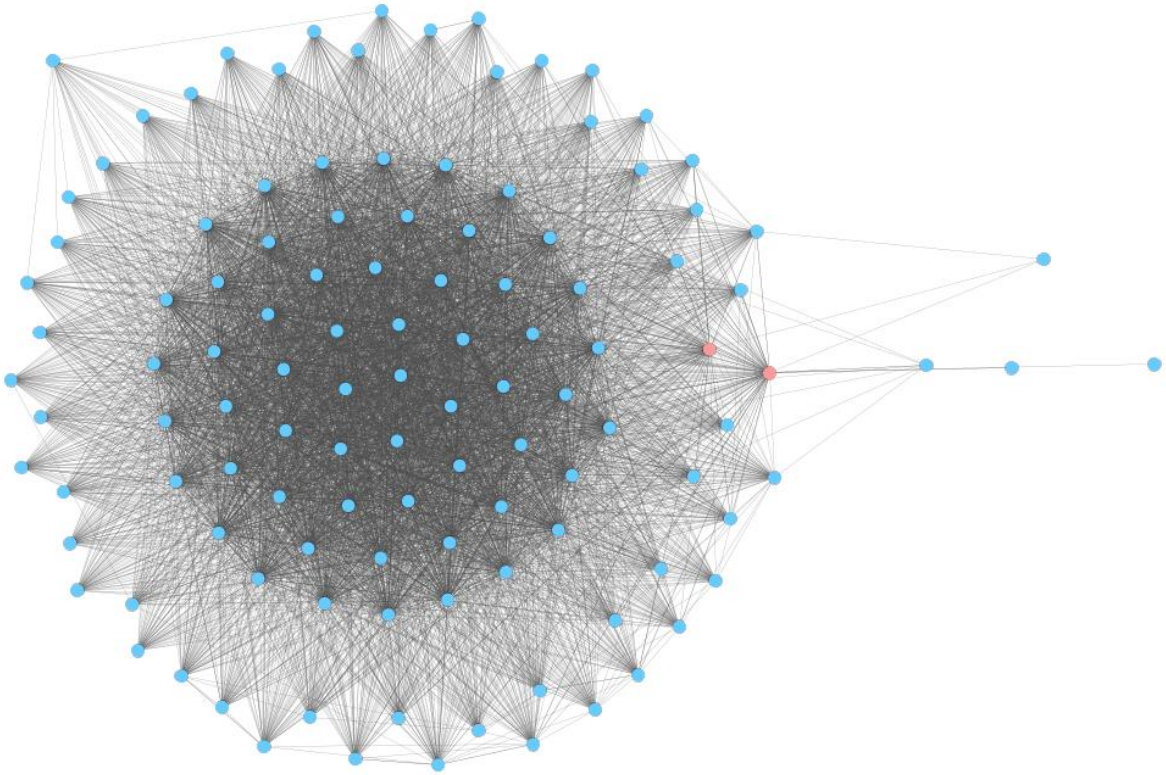

Рис. 2 Веб-граф веб-пространства компании «Газпром»

На изображении видно, что все вершины сильно связаны между собой, нет вершин, которые были бы изолированы. Исходя из полученного веб-графа было сделано предположение, что некоторые ссылки внутри веб-пространства «Газпрома» расставлены искусственно для поднятия ссылочной популярности веб-сайтов. Для проверки данного предположения была сформулирована задача оптимизации, основанная на математической модели согласованного поведения малых сообществ.

Глава 3. Решение оптимизационной задачи

В данной главе рассматриваются вопросы согласованного поведения веб-сообществ, основанные на исследовании математической модели.

3.1 Алгоритм решения

Как отмечалось ранее, будем считать, что участники Интернет-сообщества имеют потенциальную возможность согласовывать свои действия для увеличения ссылочной популярности веб-сайтов данного сообщества. Для проверки гипотезы о согласованном поведении сообщества используем математическую модель, описанную в главе 1. Алгоритм применения математической модели состоит из следующих шагов:

1. выделяется некоторое сообщество;
2. определяются значения $\hat{c}_i$ при помощи механизмов Яндекса — в качестве значений $\hat{c}_i, i = \overline{1,n}$ были взяты значения индекса качества сайта (ИКС) [30];
3. с помощью базы данных внешних ссылок, полученной при краулинге, определяются:
   - наличие или отсутствие связей между участниками сообщества,
   - значения $n,\ m_i,\ L_i,\ i = \overline{1,n}$;
4. строится реальная матрица $X^{real} = \left(x_{ij}^{real}\right),\ i,j = \overline{1,n}$;
5. вычисляется значение функционала $F(X^{real})$;
6. вычисляются исходные значения $c_i,\ i = \overline{1,n}$ на основе имеющихся значений $\hat{c}_i,\ i = \overline{1,n}$;
7. находится оптимальное решение X^{opt} для математической модели;
8. вычисляется значение функционала $F(X^{opt})$;

9. на основании отклонения значения целевой функции, вычисленной на основе реальной матрицы сообщества, от значения функции, вычисленной на матрице оптимальных значений, делается вывод о согласованном или несогласованном поведении участников сообщества.

3.2 Построение исходной матрицы

Основываясь на данных, полученных в результате работы краулера, была построена матрица X^{real}, по сути являющаяся матрицей смежности графа, вершинами которого являются участники сообщества.

Далее из матрицы были исключены строки (и соответствующие столбцы), для которых $\hat{c}_i = 0$, так как вклад данных участников не изменяет значение целевой функции математической модели. Затем, согласно ограничению $\sum_{j=1}^{n} x_{ij} = m_i\,,\ i = \overline{1,n}$, были удалены строки (и соответствующие столбцы), для которых $\sum_{j=1}^{n} x_{ij} = 0$. Процесс исключения имеет рекурсивный характер, так как удаление строк и столбцов может вести к тому, что условие $\sum_{j=1}^{n} x_{ij} = 0$ снова выполняется на уже уменьшенной матрице.

В итоге была получена матрица исследуемого Интернет-сообщества, которая является объектом дальнейших исследований на предмет согласованного поведения.

3.3 Реализация алгоритма

Исходными данными для решения оптимизационной задачи являются:

n — количество участников сообщества;

$m_1, \ldots, m_n$ — вектор прямых ссылок от каждого участника на других участников сообщества;

$L_1, \ldots, L_n$ — вектор общего числа исходящих ссылок от каждого участника;

$\hat{c}_1, \ldots, \hat{c}_n$ — вектор значений ИКС — Индекса качества сайта;

$c_1, \ldots, c_n$ — вектор “начальных” значений ИКС — Индекса качества сайта, найденный при помощи решения системы линейных уравнений.

Переменными являются элементы матрицы X.

Программная реализация алгоритма состоит из следующих шагов:

1. На основе данных, полученных из краулера строится реальная матрица X^{real}, вектор ИКС $\hat{c}_1, \ldots, \hat{c}_n$, вектор общего числа исходящих ссылок $L_1, \ldots, L_n$;
2. Отбрасываются нулевые компоненты в векторе $\hat{c}_1, \ldots, \hat{c}_n$ (а также соответствующие строки/столбцы в матрице исходящих ссылок и соответствующие компоненты вектора общего числа исходящих ссылок), так как они не изменяют значение целевой функции *Модели 1*, кроме того, они недопустимы для целевой функции *Модели 2*;
3. Из матрицы X^{real} исключаются компоненты, для которых сумма значений в строке равна 0, (см. пункт 3.2). Также отбрасываются соответствующие компоненты вектора $\hat{c}_1, \ldots, \hat{c}_n$ и вектора $L_1, \ldots, L_n$;
4. Строится оптимизированная матрица путем рекурсивного запуска процедуры, которая для каждой строки ссылочной матрицы расставляет исходное число единиц так чтобы, значение функционала для *Модели 1* (*Модели 2*) было минимальным (максимальным);
5. Находится значение функционалов от оптимизированной и реальной матриц;
6. Вычисляется итоговое значение отклонений $F1(X^{opt})/F1(X^{real})$ и $F2(X^{opt})/F2(X^{real})$.

3.4 Полученные результаты

Описанный алгоритм был реализован на языке Java. Код проекта был размещен на Github для удобства хранения и доступа к нему [31].

С помощью программной реализации алгоритма была найдена оптимальная матрица X^{opt}, а также были вычислены значения функционалов для реальной и оптимальной матриц. Полученные результаты отклонений $F1(X^{opt})/F1(X^{real})$ и $F2(X^{opt})/F2(X^{real})$ для *Моделей 1* и *2* представлены в табл. 3.

Отклонение по *Модели 1*	0.994
Отклонение по *Модели 2*	5.2

Табл. 3 Результаты расчетов

Согласно *Модели 1* поведение рассматриваемого сообщества можно трактовать как близкое к согласованному, так как значение близко к единице. Результаты *Модели 2* показывают большее отклонение матрицы ссылок от оптимума. В итоге, согласно результатам, полученным в обеих моделях, можно предположить, что в какой-то мере имело место искусственное добавление гиперссылок для увеличения ссылочной популярности веб-сайтов, входящих в состав сообщества «Газпром».

Выводы

В ходе данной работы была описана методика построения веб-графа как модели информационного веб-пространства. Предложенная методика была использована для моделирования веб-пространства крупной коммерческой компании на примере «Газпром». В результате была получена база данных внешних гиперссылок. На основе полученных данных был построен граф внешних гиперссылок веб-пространства компании «Газпром».

Также была проверена гипотеза о согласованном поведении сообщества. Для этого использовались две математические модели, имеющие различный вид оптимизационной функции. Был реализован алгоритм расчета функционалов этих моделей, при помощи которого находились матрицы оптимального размещения гиперссылок, которые затем сравнивались с реальной матрицей.

В результате, основываясь на отклонениях значений функционалов, было подтверждено предположение, сделанное на основе полученного веб-графа, о согласованных действиях веб-сайтов для увеличения ссылочной популярности.

Заключение

Вебометрические исследования играют большое значение из-за возрастающей зависимости организаций от сети Интернет. Эти исследования позволяют определить, насколько организация следит за тенденцией развития своих сайтов.

Поисковые системы играют большую роль для различных учреждений, так как от места расположения ссылок на странице поисковой выдачи зависит количество посетителей их сайтов, а значит и популярность ресурса. Так как ссылочная популярность играет не последнюю роль в механизмах ранжирования, существует возможность ее искусственного увеличения и, как следствие, повышения сайта в рейтинге поисковых систем. Поэтому, существует потенциальная возможность образования малых Интернет-сообществ, которые согласовывают свои действия для увеличения ссылочной популярности путем публикации ссылок на веб-сайты, принадлежащие сообществу. Поэтому важно уметь определять согласованные действия веб-сайтов, для предотвращения неправомерной манипуляции рейтинга.

Интернет — динамическая система. В данном работе рассматривался “снимок” на определенный момент времени, поэтому, в качестве продолжения исследований возможно периодически собирать заново данные из Интернета, чтобы анализировать, как данные в вебе и экономические показатели компаний коррелируют с течением времени.

Список литературы

1. Benoît G. Data mining // Annual Review of Information Science and Technology, 2002. Vol.36, No 1. P. 265-310.
2. Almind T., Ingwersen P. Informetric analyses on the World Wide Web: Methodological approaches to «webometrics» // Journal of Documentation, 1997. Vol. 53, No 4. P. 404-426.
3. Björneborn L. Small-world link structures across an academic web space: a library and information science approach. // Royal School of Library and Information Science, 2004. 399 p.
4. Thelwall M. Introduction to webometrics: Quantitative web research for the social sciences // Morgan & Claypool Publishers, 2009. 116 p.
5. Thelwall M. Webometrics and Social Web Research Methods // University of Wolverhampton, 2013. 142 p.
6. Esteban Romero-Frías Googling Companies - a Webometric Approach to Business Studies // Electronic Journal of Business Research Methods, 2009. Vol.7, No 1. P. 93-106.
7. Kosala R., Blockeel H. Web Mining Research: A Survey // ACM SIGKDD Explorations Newsletter, 2001. Vol. 2, No 1. P. 1-15.
8. Srivastava J., Cooley R., Deshpande M., Tan P.–N. Web Usage Mining: Discovery and
Applications of Usage Patterns from Web Data // SIGKDD Explorations, 2000. Vol. 1, No 2. P. 12-23.
9. Kumar R., Singh A.K. Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval *// American Journal of Applied Sciences,* 2010. Vol. 7, No 6. P. 840-845.
10. Liu, B. Web Data Mining // Springer. 2007. 433 p.

11. M.G. da Gomes Jr., Gong Z. Web Structure Mining: An Introduction // Proceedings of the IEEE International Conference on Information Acquisition, 2005. P. 590-595.

12. Bar-Ilan J. Data collection methods on the Web for infometric purposes: A review and analysis // Scientometrics. 2001. Vol. 50, No 1. P. 7–32.

13. Google advanced search.
https://www.google.com/advanced_search

14. Yandex advanced search.
https://yandex.ru/support/search/how-to-search/advanced-search.html

15. Yahoo advanced search.
https://search.yahoo.com/search/options?fr=fp-top&p=&guccounter=1

16. Bing advanced search.
http://help.bing.microsoft.com/#apex/18/en-us/10002/0

17. DuckDuckGo advanced search.
https://help.duckduckgo.com/duckduckgo-help-pages/results/syntax/

18. Печников А.А., Сотенко Е.М. Программы-краулеры для сбора данных о представительских сайтах заданной предметной области — аналитический разбор // Современные наукоемкие технологии, 2017. № 2. С. 58-62.

19. Borodin A., Roberts G.O., Rosental J.S., Tsaparas P. Link Analysis Ranking: Algorithms, Theory, and Experiments // ACM Transactions on Internet Technology, 2005. Vol. 5, No 1. P. 231–297.

20. Du Y., Shi Y., Zhao X. Using spam farm to boost PageRank // Proc of the 3rd International Workshop on Adversarial Information Retrieval on the Web, 2007. P. 29-36.

21. Трофименко Е.А. Оптимизация расчета ссылочной популярности и учета ее при ранжировании результатов поиска // Интернет-математика 2005. Автоматическая обработка веб-данных, 2005. С. 272-282.

22. Печников А.А. Математические модели размещения ссылок в локализованной системе Интернет-ресурсов // Системы управления и информационные технологии, 2007. № 28. С. 92-96

23. Pant G., Srinivason P., Menczer F. Crawling the Web // Web Dynamics, 2004. P. 153-177.

24. Kleinberg J.M., Kumar R., Raghavan P., Rajagopalan S., Tomkins A.S. The Web as a Graph: Measurements, Models, and Methods // Conference on Combinatorics and Computing, 1999. P. 1-18.

25. Печников А.А. Задача рационального размещения ссылок в регламентируемой локализованной системе Интернет-ресурсов // Труды Института прикладных математических исследований КарНЦ РАН, 2006. № 7. С. 176-182.

26. Печников А.А. Математические модели размещения ссылок в локализованной системе интернет-ресурсов // Системы управления и информационные технологии, 2007. № 28. С. 92-96.

27. GitHub – yasserg/crawler4j: Open Source Web Crawler for Java. https://github.com/yasserg/crawler4j

28. PostgreSQL. https://www.postgresql.org/

29. Gephi. https://gephi.org/

30. Индекс качества сайта. https://yandex.ru/support/webmaster/site-quality-index.html

31. Репозиторий разработанного проекта. https://github.com/Suisel/Crawler